Development of Brief Rating Scales for Progress Monitoring Internalizing Behavior

Amy M. Briesch[1], Aberdine R. Donaldson[1], Michael Matta[2], Robert J. Volpe[1], Brian Daniels[3], Julie Sarno Owens[4]

[1]Department of Applied Psychology, Northeastern University

[2]Department of Psychological, Health, and Learning Sciences, University of Houston

[3]Department of Counseling and School Psychology, University of Massachusetts, Boston

[4]Department of Psychology, Ohio University

Journal of Emotional and Behavioral Disorders, Published online August 23, 2021

Author Note

Correspondence concerning this article should be addressed to Amy M. Briesch at 404 International Village, Northeastern University, Boston MA 02115

Acknowledgement: Preparation of this article was supported through funding by the Institute for Education Sciences, U.S. Department of Education (R324A150071). Opinions expressed herein do not necessarily reflect the position of the U.S. Department of Education, and such endorsements should not be inferred.

Abstract

Prevalence estimates suggest that up to 20% of students in schools experience significant internalizing behaviors that impact behavioral, social or academic functioning. School-based interventions have great potential to promote student mental health; however, validated and feasible brief assessments are needed to progress monitor students' response to these supports. The purpose of the current study was two-fold: to (a) develop and validate teacher-completed brief rating scales for progress monitoring internalizing concerns in elementary-aged students and (b) to determine the reliability of the resultant measures. First, item content was generated and subjected to evaluation by two panels of school-based consumers and researchers. Within the second phase of development, exploratory and confirmatory factor analyses were used to reduce the initial number of items and ensure that the items were indicators of one latent factor. Teachers in grades K through 3rd (N = 307) each completed ratings for one randomly-selected student in their classroom. Results of factor analysis for each scale indicated one-factor solutions for the 4-item Anxious/Depressed (ω = .88) and 4-item Socially Withdrawn (ω = .87) scales.

*Keywords***:** brief rating scale; progress monitoring; behavioral assessment; internalizing problems; exploratory factor analysis

Development of Brief Rating Scales for Progress Monitoring Internalizing Behavior

Results of the 2016 National Survey of Children's Health found that 3% of youth aged 3-17 currently struggle with depression and as many as 7% struggle with anxiety in the United State (Ghandour et al., 2019). Unfortunately, anxiety and depression—like other internalizing behaviors (e.g., social withdrawal, somatic complaints)—are often overlooked in schools, given the fact that these behaviors tend to be inner-directed and over-controlled (Sanders et al., 1999). Although internalizing behaviors may have less of an impact on the classroom environment than outwardly-directed externalizing behaviors (e.g., aggression, hyperactivity), the consequences for these students are considerable. In addition to impacting peer relations (Flook et al., 2005), research has found internalizing problems to be significantly related to both lower academic achievement (Rapport et al., 2001) and rates of high school completion (Duchesne et al., 2008).

In the last decade, several examples of effective school-based prevention and intervention programs targeting internalizing behaviors have been documented in the literature (Calear & Christensen, 2010; Farahmand et al., 2011; Herzig-Anderson et al., 2012; Neil & Christensen, 2009). This research has resulted in an expanded toolbox for the school-based practitioner with regard to effective yet feasible approaches for both preventing and reducing internalizing concerns. Consistent with a multi-tiered approach to school-based service delivery, programs exist at the universal (e.g., FRIENDS for Life; Barrett et al., 2000), secondary (e.g., Cool Kids; Schniering et al., 2006), and tertiary (e.g., Coping Cat; Kendall & Hedtke, 2006) levels for supporting different levels of mental health need. Unfortunately, however, there has been a lag in the development of appropriate assessment tools for determining the effectiveness of interventions targeting internalizing behaviors within multi-tiered systems of support (MTSS).

Assessing Internalizing Behavior within Multi-Tiered Systems of Support

The success of MTSS relies in no small part on the availability of appropriate progress monitoring tools (Center on Response to Intervention, n.d.). Progress monitoring refers to the practice of collecting and analyzing information about student behavior that can be used to evaluate the effectiveness of an intervention and measure progress towards student goals (National Center on Progress Monitoring, 2006). Dart et al. (2019) recently conducted a systematic review of the literature in order to identify tools for progress monitoring internalizing behaviors that were both psychometrically sound and feasible for use in school settings. Their review returned a total of 15 unique assessment tools, over half of which (i.e. 53%) were rating scales completed by students themselves.

Heavy reliance on self-report in the assessment of internalizing problems is not surprising given the nature of these behaviors. Behaviors such as sadness, worry, and fear are less readily observable than externalizing concerns and therefore have been shown to be more difficult for teachers to recognize (Merrell & Gueldner, 2010) and less likely for teachers to refer (e.g., Bradshaw et al., 2008), particularly when behaviors are less severe (e.g., Splett et al., 2019). Combined with the fact that teachers can only provide their perception of the child's inner experience (Smith, 2007), it is commonly argued that the best informant in the assessment of internalizing concerns is the student him or herself (Merrell, 2008). At the same time, however, the challenges of using self-report with young children are well known. For example, student ratings may be heavily influenced by a social desirability bias, or the desire to present one's self in a positive light (Harter, 1986). Young students may also not be fully aware of their impairments in functioning, particularly when those impairments involve social interactions with peers (Webster-Stratton et al. 2004). For these reasons, experts generally discourage the use of self-report with children below the age of 8 or 9 (Levitt & Merrell, 2009).

In the absence of the use of self-report to assess internalizing behaviors in young children, one alternative involves the use of teacher report. To date, however, there have been limited applications of teacher report to the assessment of internalizing concerns, with the recent review by Dart et al. (2019) identifying only two such tools: daily point sheets (Puddy et al., 2008) and Direct Behavior Rating (DBR; e.g., Dart et al., 2015; von der Embse et al., 2015). In the one study to examine the psychometric properties of daily point sheets, Puddy and colleagues (2008) had adult raters (i.e. clinicians, parents, teachers) use a 5-point scale (i.e. very poor to excellent) to assess 46 students enrolled in a school-based Intensive Mental Health Program every half hour across three to four internalizing behaviors. Although the authors found a significant correlation (i.e. .41) between the number of points earned and scores from the Child and Adolescent Functional Assessment Scale (CAFAS; Hodges & Wong, 1996), it is notable that the authors computed the mean percentage of points earned across a four-week period. As such, the psychometric properties of the measure when used daily or weekly are unknown. More recently, in seeking to evaluate the effectiveness of a peer-mediated intervention in reducing internalizing behaviors, Dart et al. (2015) measured the behavior of three elementary school students using a DBR-Multi-Item Scale (DBR-MIS). The authors created an idiographic 4-point scale for each student participant based on the results of a school-wide screening measure. Although graphed data demonstrated that each DBR-MIS was sensitive to changes in student behavior as a function of the CICO intervention, the reliability and validity of the resultant data were not specifically investigated. Thus, additional research involving school-age youth is warranted.

Both daily point sheets and DBR involve brief ratings of target behaviors that are completed by teachers at the end of a pre-determined observation period. Akin to behavior rating

scales, teachers typically use a Likert-type scale to estimate the frequency or intensity of a student's behavior; however, both types of ratings are conducted in closer temporal proximity to the actual occurrence of behavior (e.g., at the end of an instructional period as opposed to reflecting on behavior over a longer period of time).

Purpose of the Present Study

Use of a multi-informant assessment approach is commonly recommended and utilized in school-based practice, given that each informant may contribute unique information needed to fully understand the student's functioning across settings (e.g., Achenbach et al., 1986; De Los Reyes et al., 2019). When combined with concerns regarding the reliability and accuracy of self-report data in young children (Merrell, 2008; Webster-Stratton et al., 2004), it appears important to consider how teacher-completed measures may be feasibly incorporated into the school-based assessment of student internalizing behaviors. To date, only two studies have examined the use of teacher ratings to progress monitor student internalizing behaviors (e.g., Dart et al., 2015; Puddy et al., 2008); however, psychometric evidence was largely anecdotal. Therefore, the purpose of the current study was to develop and assess the psychometric properties of teacher-completed brief rating scales for progress monitoring internalizing behaviors in early elementary-aged students. Given the use of teachers as informants, the goal was to focus on those observable behaviors that correspond with internalizing concerns (Merrell & Gueldner, 2010). Proposed item content was first evaluated by a panel of school-based consumers and a panel of researchers. The construct validity of the two proposed scales (i.e., Socially Withdrawn, Anxious-Depressed) was then subsequently evaluated through exploratory and confirmatory factor analyses. In addition, the current study aimed to assess the internal consistency reliability of the two scales and whether the number of response categories was appropriate.

Method

Participants

Participants included a total of 307 K-3 teachers from 13 states and 35 school districts. Most were female teachers (96%) located in the Northeastern United States (73%). Efforts were made to recruit similar proportions of teachers at each grade level, with 26.7% in kindergarten, 20.8% in first grade, 25.7% in second grade, and 26.7% in third grade. Participants were recruited by sending emails to building principals and school psychologists within the principal investigators' partner networks asking them to serve as local coordinators for data collection. In turn, interested principals and school psychologists distributed information to general and special education teachers in their school buildings regarding the purpose and procedures of the study.

Each teacher participant was asked to complete ratings for one student in their classroom (see Procedures for additional detail). The majority of students (60.9%) were male and approximately 36% of the students were receiving special education services through an Individualized Education Programs (IEP). The sample of students included the following racial/ethnic composition: White = 67.1%, Black = 13.0%, Latino = 15.0%, Asian = 3.3%, American/Alaska Native = 1.0%, Hawaiian/Pacific Islander = 0.3%, and Unknown = 5.2%.

Procedures

To maximize both feasibility and psychometric adequacy, our goal in the current study was to develop brief rating scales with no more than five items. The first phase of the study involved item generation and refinement. Once an initial pool of items was established, a series of exploratory factor analyses was then used to reduce the overall length of the two scales, and a series of confirmatory factor analyses was conducted to test the goodness of fit of the suggested

structures. Finally, we measured reliability and plotted the item characteristic curves to inspect whether the number of response categories for each item was adequate.

Item Development

The initial pool of potential items was developed and refined in two phases. Within the first phase, we conducted an extensive search of existing measures (e.g., rating scales, observation codes) that assessed internalizing behaviors. Potential measures were identified by reviewing Buros Mental Measurements Yearbook, test publishers' catalogs, and intervention studies. Although there are four broad categories of internalizing behaviors (i.e., anxiety, depression, somatic complaints, social withdrawal; Levitt & Merrell, 2009), somatic complaints were not prioritized within the current study given the limited perceived utility by school-based stakeholders of such a scale within a progress monitoring context. As such, the keywords used to guide these searches included "anxi*," "depress*," "withdraw*," and "internalizing." Any measure that appeared to assess internalizing behaviors in school-age youth was obtained and the relevant item(s) entered into a spreadsheet.

Once all items had been entered, they were reviewed in two stages by members of the research team. First, all items were independently rated by three members of the research team on their observability using a 0-4 scale. Given that the goal was to create scales that would be completed by a classroom teacher, any items that were deemed not observable (e.g., assessed internal states such as thoughts or feelings) were excluded. Whereas a total of 28 items representing Socially Withdrawn behavior were initially generated, item content for the Anxious and Depressed scales was comparatively very limited given the focus on observable behaviors (i.e., 10 items total). As such, these items were combined to represent one potential Anxious/Depressed scale.

Within the second phase, after the research team reviewed the initial item pool and deleted any redundancies (i.e., items assessing the same behavior), feedback regarding the 15 items (i.e., six Socially Withdrawn, nine Anxious/Depressed) was next obtained from two panels of stakeholders. Members of the Consumer Advisory Panel (CAP) were recruited from partner elementary schools and were purposively selected to represent urban, suburban, and rural school districts. Initial outreach was made to building principals, who were then either invited to participate themselves or asked to nominate a psychologist, teacher, or parent. The CAP ultimately consisted of two principals, two school psychologists, four teachers, and four parents of students in grades K-3. Given that the information solicited from CAP members was not sensitive in nature and data were deidentified, the study was deemed exempt from review by the University HSIRB. Members of the Scientific Advisory Panel (SAP) were purposively recruited for their content expertise. These five researchers possessed collective expertise in the areas of scale development, formative assessment, and the constructs of interest.

For each item, members used a 5-point Likert-type scale (0 = “Strongly Disagree;” 4 = “Strongly Agree”) to evaluate the degree to which the item represented a behavior that (a) a teacher could readily see in a classroom (i.e., observability), (b) would be a suitable target for intervention (i.e., malleability), and (c) if changed would be helpful to the student and/or the classroom environment (i.e., socially valid). Additionally, members of the SAP evaluated the degree to which the item was believed to be a strong indicator of the construct to which it was assigned (i.e., construct validity). Both panels also provided feedback regarding the clarity of the items, and whether additional behavioral targets should be included. Two items were subsequently deleted from the Anxious/Depressed scale due to low scores received from the

CAP and SAP (see Table 1). This resulted in six Socially Withdrawn and seven Anxious/Depressed items that were utilized in the factor analytic studies.

Scale Refinement

The 13 items developed to measure internalizing problems (Anxious/Depressed = 7; Socially Withdrawn = 6) within the first phase of this study were subsequently administered to the 307 K-3 teachers previously described. Items were rated using a 7-point Likert-type scale, which asked the teacher to rate the degree to which they believed the behavior to be a problem (i.e., 0 = "Not a Problem," 6 = "Serious Problem") over the course of the school week. A 7-point scale was used given prior research findings that at least seven scale gradients are needed to detect small changes in behavior over time (Christ et al., 2009). Although traditional rating scales often ask teachers to rate the frequency with which a behavior occurred (e.g., "Never" to "Always"), one of the problems with such a scaling approach is that it does not take into account the intensity of the behavior. For example, two students may express worries "frequently;" however, one student may engage in catastrophic worrying that substantially impairs their functioning whereas another student's worries may be less impactful on a daily basis. As such, asking teachers to rate the extent to which they believed a behavior to be a problem was designed to encompass both frequency and severity/impairment.

The research team assigned each of the teacher participant a random number between 1 and 20 using a random number generator and instructed the teacher to rate the student corresponding to that number on his or her alphabetical class list. In the case that the provided number was higher than the number of students in their class, the teacher was instructed to continue counting from the last student on their list. This approach was used to ensure sufficient variability with regard to the endorsement of student concerns. Additionally, teachers were asked

to provide feedback concerning the clarity of the instructions, items, and response scale. All ratings and feedback were completed electronically within the Qualtrics platform. Because the study was deemed to present minimal risk of harm to teacher respondents and identifying information was not collected for student participants, the project was deemed exempt from review by the University Institutional Review Board. All respondents received an information sheet outlining the purpose and procedures of the study before making the decision to proceed with the online survey.

Data Analysis

The analyses were conducted in RStudio 1.4 (RStudio Team, 2020). No missing data were expected due to the procedures of this study, which required teachers to rate every item of the scales prior to submitting responses electronically. The dataset was randomly split into two subsets; we conducted exploratory factor analyses (EFA) for item reduction purposes on the first subset and confirmatory factor analyses (CFA) to validate the factor structure on the second subset. Due to the nature of the constructs examined, we expected the majority of the items to be asymmetrical. Although the use of five or more response categories generally allows to approximate Likert-type items to continuous variables, estimates are likely to be biased when dealing with skewed and highly kurtotic distributions (Rhemtulla et al., 2012). Therefore, we conducted EFA using the weighted least squares (WLS) estimator and CFA using the robust WLS mean and variance adjusted (WLSMVS) approaches, which do not require the observed variables to be continuous or to follow normal distributions (Kline, 2016).

First, we conducted the Kaiser-Meyer-Olkin (KMO) test and the Bartlett's Test of Sphericity to ensure that the data were deemed appropriate for factor analysis (McCoach et al., 2013). We then used EFA on each scale individually to select a maximum of five items with

strong loadings on the latent factor. EFA was conducted via the psych R package (Revelle, 2021) using polychoric correlations which assume that ordinal items are dichotomized versions of continuous, normally distributed latent variables and give unbiased coefficients. Regarding the item loadings, coefficients below .45 were interpreted as poor indicators of the latent factor (Comrey & Lee, 1992). In addition, we inspected the communality coefficients (h^2) to avoid retaining items with a low degree of shared variance, which could result in the derivation of a set of meaningless factors (Pett et al., 2003). Communalities above .40 are generally considered desirable and indicate that items cluster closely together with the other items (Fabrigar et al., 1999).

Second, CFA was performed on the other half of the sample using the lavaan R package (Rosseel et al., 2012) to validate the factor model obtained from the exploratory analyses. The confirmatory models were estimated with the cumulative probit link function and the theta parameterization; this implies that the residuals of the latent response propensity are assumed to be normally distributed with a mean of 0 and standard deviation of 1. The models were assessed through a variety of fit indices which evaluate different aspects of misspecification. Fit indices include the Comparative Fit Index (CFI), the Tucker–Lewis Index (TLI), the Root Mean Square Error of Approximation (RMSEA), and the Standardized Root Mean Square Residual (SRMSR). Fit indices were interpreted based on rules of thumbs accepted by the scientific community. Good-fitting models have been often associated with CFI and TLI greater than .90 or .95 and RMSEA and SRMR less than .08 (Hu & Bentler, 1999; McDonald & Ho, 2002). However, simulation studies and empirical evidence have shown that these indices depend on many factors of the models, such as the magnitude of the loadings and the sample size (Kline, 2015). It is also important to note that fit indices are rarely used in isolation and often lead to different

conclusions when used in the context of categorical or ordinal data. Therefore, the degree of "reasonableness" associated with a model requires an overall evaluation of the fit indices rather than rigid applications of cut-off values.

Finally, we calculated internal consistency and item characteristic curves using the full sample. The internal consistency of the scales was calculated with omega coefficients via the semTools R package (Jorgensen et al., 2021). The item characteristic curves (ICC) use the coefficients estimated for confirmatory models and show the probability of endorsing the levels of the Likert scale for different levels of the latent response propensity. The curves were plotted with ggplot2 (Wickham, 2016).

Results

The Kaiser-Meyer-Oklin (KMO) test and Bartlett's Test of Sphericity conducted on the two scales separately indicated that the variables could be analyzed through a factor analytic approach. The KMO statistics exceeded the suggested value of .50 (Hair et al., 2006), and the Bartlett's Test of Sphericity reached statistical significance ($p < .05$). Based on these indicators, the correlation matrices were deemed appropriate for factor analyses. As was expected, the distribution of several items composing the two scales was either skewed or highly kurtotic (see Table 2). This finding validated the use of WLS estimators for the subsequent analyses.

Exploratory Factor Analysis

The EFA performed on the polychoric correlation matrix of the six Anxious-Depressed items revealed the presence of one latent factor, which was consistent with inspection of the scree plot and the results of parallel analysis. One factor was therefore extracted, which explained 46.00% of the observed variance. All items demonstrated factor loadings above .40 (see Table 4); however, three items (i.e., Sick, Cries, and Irritable) were eliminated due to low

communality coefficients (i.e., < .40), suggesting that these items were not well represented in the factor.

The EFA performed on the polychoric correlation matrix of the five Socially Withdrawn items also revealed the presence of one latent factor, which explained 50.00% of the variance. This was consistent with results of parallel analysis as well as inspection of the scree plot, which demonstrated a clear break after the first eigenvalue. All items demonstrated factor loadings above .40 (see Table 2). Two items (i.e., Quiet and Shy) had communality coefficients below .20; hence, these were deleted from the scale and the subsequent analyses.

Confirmatory Factory Analysis

Next, we tested the one-factor model using CFA for each scale separately. Four indicators were included in the Anxious Depressed model (Table 3). All the items showed strong loadings on the latent factor ranging from .79 (Sad) to .90 (Worries). Model fit results were good for Anxious Depressed items (CFI = 0.995, TLI = 0.992, SRMR = 0.019). It should not be surprising that the RMSEA indicated poor fit (RMSEA = 0.128, 90% CI [0.000, 0.344]) in that this index tends to be biased in models with small degrees of freedom and estimated for small samples as a consequence of the bias affecting the χ^2 (Kenny et al., 2015; Shi et al., 2019). However, the inspection of the residual correlation matrix showed that no coefficients were above .10 so no major misspecifications affected the model. In other words, the observed correlation matrix was well represented by the implied matrix (Kline, 2015).

Four indicators were included in the Socially Withdrawn model (Table 4). All of the items showed strong loadings on the latent factor ranging from .83 (Does not join) to .94 (Avoid). Model fit results were good also for Socially Withdrawn items (CFI = 0.990, TLI = 0.983, SRMR = 0.034). Again, the RMSEA indicated poor fit (RMSEA 0.181, 90% CI [0.092,

0.292]). However, the inspection of the residual correlation matrix revealed that no coefficients were above .10 so no major misspecifications affected the implied model.

Reliability Analysis

After determining the goodness of fit, we tested the same models on the full sample via CFA. We obtained similar results with RMSEA above the cut-off of .08, but no indication of model misspecification from the residual correlation matrix. Then, we used the full sample to calculate the internal consistency for each of the two internalizing scales. McDonald's Omega (ω) was found to be strong for the Anxious-Depressed scale (ω = .87), as well as for the Socially Withdrawn scale (ω = .88).

Item Characteristic Curves

Fig. 1 and 2 show the ICC for the items included in the confirmatory models for the two scales. Each plot displays the probability to respond in different categories at different levels of the latent response propensity. The number of curves corresponds to the number of categories of the Likert scale. Two patterns can be noted across the items. First, two or more curves within each plot were generally much lower than all the other curves; this means that teachers never used the corresponding category on the Likert scale more frequently at specific levels of the latent construct. This finding suggests that those categories might not be particularly useful to discriminate among students. Second, all the response categories except the first one reached the peak at the right of 0, which represents the average factor score in the sample. This means that there is a high probability that a person with an average response propensity is going to respond in the first category (i.e., Not a Problem). Obviously, there is still a low percentage that the person will respond in higher categories, but it will take very high levels of the latent construct to make the teacher endorse the items past the first or second category.

Discussion

The success of MTSS relies on the availability of psychometrically sound—yet feasible—progress monitoring measures for assessing students' response to intervention. Given that as many as one in five students in schools may experience impairment as the result of internalizing behaviors (Walker et al., 2000), there exists a particular need for psychometrically defensible tools for monitoring student responsiveness in the internalizing domain. The goal of the current paper was to describe the development and initial validation of two teacher-completed brief rating scales specifically designed to progress monitor Anxious/Depressed and Socially Withdrawn behaviors.

The challenge in creating a brief rating scale that could be used by classroom teachers to progress monitor internalizing behaviors was the fact that the items had to be observable. For this reason, all items were rated by Consumer- and Scientific Advisory Panels concerning their observability and any items believed to not be appropriate for teacher observation were omitted from the scales. In order to reduce the pool of generated items and to ensure that the final items represented a unitary, reliable construct, EFA, CFA, and reliability analyses were subsequently conducted for each scale. Ultimately, strong internal consistency was found for both Anxious-Depressed and Socially Withdrawn scales consisting of four items.

Both of these newly developed brief rating scales address internalizing concerns that may impact students both in school and beyond. The Socially Withdrawn scale was designed to measure the degree to which a student isolates themself from the larger peer group. In addition to experiencing rejection from peers and poor-quality relationships while in school, there are also several potential long-term consequences for students who are socially withdrawn in childhood (Rubin et al., 2009). A study by Rubin and colleagues (1995), for example, found that students

who were socially withdrawn at age 7 were more likely to experience higher levels of loneliness and depression, as well as lower levels of self-esteem, at age 14. The Anxious-Depressed scale, on the other hand, was designed to assess observable behaviors that are indicative of negative mood or anxiety. Studies have shown anxiety and depression to be related to a number of poor outcomes in youth including impaired social relations, poor self-esteem, and substance abuse (Farrell & Barrett, 2003; McLoone et al., 2006). Early intervention for students struggling with internalizing concerns is therefore important, and results of the current study suggest that a brief 4-item rating scale may provide a reliable estimate of these behaviors without overtaxing respondents in the school setting.

Limitations

Although the results of the current study provide initial psychometric evidence in support of two brief rating scales to assess internalizing behavior, limitations of this work must be acknowledged. First, the sample utilized within the current study was one of convenience. Of note was the fact that the majority of student participants came from the Northeastern United States. In addition, limited demographic information was collected about the teacher and student participants. Thus, it is difficult to draw conclusions regarding the generalizability of the obtained psychometric findings.

Second, it was notable that the mean scores across all items were fairly low. In fact, the highest mean rating was only 1.62 on a 7-point Likert scale. The ICCs also indicated that overall teachers were more likely to indicate that a behavior was “not a problem” and did not use the full range of gradients on the scales. This is not surprising, given that teachers were asked to randomly select one student in their classroom. The reported rates of internalizing disorders in young children are low, with prevalence estimates for children under the age of 11 reported as

low as 3% (Cartwright-Hatton et al., 2006). In addition, it is also possible that despite efforts to select those manifestations of internalizing problems that were outwardly observable, it was still challenging for teachers to observe the target behaviors within the classroom setting. Targeting use of these scales with those students referred for internalizing concerns across a wider range of settings (e.g., playground, lunchroom) will be an important next step for future research.

Implications for Practice and Directions for Future Research

In recent years, increased attention has been paid to ensuring representation of internalizing concerns within school-based screening measures (e.g., Social, Academic, & Emotional Behavior Risk Screener: Kilgus et al., 2013; Student Risk Screening Scale-Internalizing and Externalizing: Lane et al., 2012). Proactive screening efforts are critical to ensure that those students who are often overlooked in the school setting are identified early and provided with the appropriate mental health supports. The missing link, however, has been the availability of feasible yet psychometrically defensible tools for monitoring student response to school-based supports once provided. Results of the current study begin to lay the foundation for the psychometric evidence base for brief rating scales targeting internalizing concerns by demonstrating that both scales reliably measure unitary constructs at one point in time. There are, however, additional questions that need to be answered before these scales can be fully endorsed as progress monitoring tools and adopted by practitioners within their local practice. As outlined by the National Center on Intensive Intervention (n.d.), there are several quality indicators that together help to establish the technical rigor of behavioral progress monitoring tools. Perhaps the most important criterion within a progress monitoring context is that the tools must demonstrate sensitivity to change in response to intervention. Future research is therefore needed to document that meaningful changes in brief rating scale scores would be observed as a function of

implementing evidence-based intervention within the school setting. Furthermore, work is needed to understand how many ratings are needed to obtain a dependable estimate of these two classes of internalizing behavior.

In addition, given the discrepancies that have been documented between teacher and student ratings (Achenbach et al., 1987), it will be important to explore the degree to which teacher- and student-completed brief rating scales of internalizing concerns correspond with one another. One of the reasons why use of self-report has generally been avoided when working with younger children is that the reliability of such reporting has been questioned. Specifically, it has been suggested that children in the early elementary grades are more likely to rate themselves as they would like to be seen (i.e. “ideal self”) as opposed to how they currently are (“real self”; Harter, 1986). Much of this research, however, has been conducted using traditional rating scales, in which respondents are asked to reflect on behavior over a period of two weeks or more. It is therefore unknown how the accuracy of student ratings might improve given a shorter rating period (e.g., a single day).

References

Achenbach, T. M., McConaughy, S. H., & Howell, C. T. (1987). Child/adolescent behavioral and emotional problems: Implications of cross-informant correlations for situational specificity. *Psychological Bulletin, 101,* 213-232.

Barrett, P. M., Webster, H., & Turner, C. (2000). *FRIENDS prevention of anxiety and depression for children group leader's manual*. Australian Academic Press.

Beck, J. S., Beck, A. T., & Jolly, J. B. (2001). *Beck Youth Inventories*. Psychological Corporation.

Bradshaw, C. P., Buckley, J. A., & Ialongo, N. S. (2008). School-based service utilization among urban children with early onset educational and mental health problems: The squeaky wheel phenomenon. *School Psychology Quarterly, 23*, 169-186.

Briesch, A. M., Chafouleas, S. M., & Riley-Tillman, T. C. (2016). *Direct Behavior Ratings: Linking Assessment, Communication, and Intervention*. Guilford Press.

Calear, A. L., & Christensen, H. (2010). Systematic review of school-based prevention and early intervention programs for depression. *Journal of Adolescence, 33*, 429-438.

Cartwright-Hatton, S., McNicol, K., & Doubleday, E. (2006). Anxiety in a neglected population: Prevalence of anxiety disorders in pre-adolescent children. *Clinical Psychology Review, 26*, 817-833.

Christ, T. J., Riley-Tillman, T. C., & Chafouleas, S. M. (2009). Foundation for the development and use of Direct Behavior Rating (DBR) to assess and evaluate student behavior. *Assessment for Effective Intervention, 34,* 201–213. doi:10.1177/1534508409340390

Comrey, A. L., & Lee, H. B. (1992). *A first course in factor analysis* (2nd ed.). Erlbaum.

Cook, C. R., Volpe, R. J., & Delport, J. (2014). Systematic progress monitoring of students with

emotional and behavioral disorders. In H. M. Walker, & F. M. Gresham (Eds.). *Handbook of evidence-based practices for emotional and behavioral disorders: Applications in schools* (pp. 211–228). Guilford Press.

Crone, D. A., Hawken, L. S., & Horner, R. H. (2010). *Responding to problem behavior in schools: The Behavior Education Program (2nd Ed.).* Guilford Press.

Cunningham, J. M., & Suldo, S. M. (2014). Accuracy of teachers in identifying elementary school students who report at-risk levels of anxiety and depression. *School Mental Health, 6,* 237-250.

Daniels, B., Volpe, R. J., Briesch, A. M., & Gadow, K. D. (2017). Dependability and treatment sensitivity of multi-item direct behavior rating scales for interpersonal peer conflict. *Assessment for Effective Intervention, 43,* 48-59

Dart, E. H., Arora, P. G., Collins, T. A., & Doll, B. (2019). Progress monitoring measures for internalizing symptoms: A systematic review of the peer-reviewed literature. *School Mental Health, 11,* 265–275.

Dart, E. H., Furlow, C. M., Collins, T. A., Brewer, E., Gresham, F. M., & Chenier, K. H. (2015). Peer-mediated check-in/check-out for students at-risk for internalizing disorders. *School Psychology Quarterly*, *30*(2), 229-243. doi:10.1037/spq0000092

De Los Reyes, A., Cook, C. R., Gresham, F. M., Makol, B. A., & Wang, M. (2019). Informant discrepancies in assessments of psychosocial functioning in school-based services and research: Review and directions for future research. *Journal of School Psychology*, *74*, 74-89.

Duchesne, S., Vitaro, F., Larose, S., & Tremblay, R. E. (2008). Trajectories of anxiety during elementary-school years and the prediction of high school noncompletion. *Journal of*

Youth and Adolescence, 37, 1134-1146.

Evans, S. W., & Youngstrom, E. (2006). Evidence-based assessment of Attention-Deficit Hyperactivity Disorder. *Journal of the American Academy of Child and Adolescent Psychiatry, 45,* 1132-1137.

Fabiano, G. A., Vujnovic, R. K., Pelham, W. E. Waschbusch, D. A., Massetti, G. M., Pariseau, M. E., & Volker, M. (2010). Enhancing the effectiveness of special education programming for children with Attention Deficit Hyperactivity Disorder using a Daily Report Card. *School Psychology Review, 39,* 219-239.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods, 4,* 272–299.

Farahmand, F. K., Grant, K. E., Polo, A. J., & Duffy, S. N. (2011). School-based mental health and behavioral programs for low-income, urban youth: A systematic and meta-analytic review. *Clinical Psychology: Science and Practice, 18,* 372-390.

Farrell, L. J., & Barrett, P. M. (2003). Prevention of childhood emotional disorders: Reducing the burden of suffering associated with anxiety and depression. *Child and Adolescent Mental Health, 12,* 58–65.

Flook, L., Repetti, R. L., & Ullman, J. B. (2005). Classroom social experiences as predictors of academic performance. *Developmental Psychology, 41,* 319-327.

Ghandour, R. M., Sherman, L. J., Vladutiu, C. J., Ali, M. M., Lynch, S. E., Bitsko, R. H., & Blumberg, S. J. (2019). Prevalence and treatment of depression, anxiety, and conduct problems in U.S. children. *The Journal of Pediatrics,* 206, 256-267. https://doi.org/10.1016/j.jpeds.2018.09.021.

Gresham, F. M., Cook, C. R., Collins, T., Dart, E., Rasetshwane, K., Truelson, E., & Grant, S. (2010). Developing a change-sensitive brief behavior rating scale as a progress monitoring tool for social behavior: An example using the Social Skills Rating System-Teacher Form. *School Psychology Review, 39*, 364-379.

Hair, J. F., Jr, Black, B., Babin, B. J., Anderson, R. E., & Tatham, R. L. (2006). *Multivariate Data analysis (Sixth edition)*. Pearson Prentice Hall.

Harter, S. (1986). Processes underlying the construction, maintenance, and enhancement of self-concept in children. In J. Suis & A. Greenwald (Eds.), *Psychological perspectives on the self.* (pp. 137-181). Erlbaum.

Herzig-Anderson, K., Colognori, D., Fox, J. K., Stewart, C. E., & Warner, C. M. (2012). School-based anxiety treatments for children and adolescents. *Child and Adolescent Psychiatric Clinics of North America, 21*, 655–668. http://doi.org/10.1016/j.chc.2012.05.006

Hodges, K., & Wong, M. M. (1996). Psychometric characteristics of a multidimensional measure to assess impairment: The Child and Adolescent Functional Assessment Scale (CAFAS). *Journal of Child and Family Studies, 5,* 445–467.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives, *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55, http://doi.org/10.1080/10705519909540118

Jorgensen, T. D., Pornprasertmanit, S., Schoemann, A. M., & Rosseel, Y. (2021). semTools: Useful tools for structural equation modeling. R package version 0.5-4 [Computer software]. Retrieved from https://CRAN.R-project.org/package=semTools

Kendall, P. C., & Hedtke, K. (2006). *Coping Cat workbook. (2nd ed)*. Workbook Publishing.

Kenny, D. A., Kaniskan, B., & McCoach, D. B. (2015). The performance of RMSEA in models

with small degrees of freedom. *Sociological Methods & Research*, *44*(3), 486–507. https://doi.org/10.1177/0049124114543236

Kline, R. B. (2016). *Principles and practice of structural equation modeling*. Guilford publications.

Kovacs, M. (1992). *Children's Depression Inventory*. Multi-Health Systems.

Lane, K. L., Oakes, W. P., Harris, P. J., Menzies, H. M., Cox, M., & Lambert, W. (2012). Initial evidence for the reliability and validity of the student risk screening scale for internalizing and externalizing behaviors at the elementary level. *Behavioral Disorders, 37*, 99-122.

Levitt, V. H., & Merrell, K. W. (2009). Linking assessment to intervention for internalizing problems of children and adolescents. *School Psychology Forum, 3*, 13-26.

McCoach, D. B., Gable, R. K., & Madura, J. (2013). *Instrument design in the affective domain*. (Third Edition). Springer.

McDonald, R. P., & Ho, M.-H. R. (2002). Principles and practice in reporting structural equation analyses. *Psychological Methods*, *7*(1), 64–82. https://doi.org/10.1037/1082-989X.7.1.64

McLoone, J., Hudson, J. L., & Rapee, R. M. (2006). Treating anxiety disorders in a school setting. *Education and Treatment of Children, 29*, 219–242.

Merrell, K.W. (2008). *Helping students overcome depression and anxiety: A practical guide*. Guilford Press.

Merrell, K. W., & Gueldner, B. A. (2010). Preventive Interventions for students with internalizing disorders: Effective strategies for promoting mental health in schools. In M. R. Shinn & H. M. Walker (Authors), *Interventions for achievement and behavior*

problems in a three-Tier model including RTI. National Association of School Psychologists.

Miller, F. G., Riley-Tillman, T. C., & Chafouleas, S. M. (2016). Use of DBR in progress monitoring (pp. 78-98). In A. M. Briesch, S. M. Chafouleas, & T. C. Riley-Tillman (Eds), *Direct Behavior Ratings: Linking Assessment, Communication, and Intervention*. Guilford Press.

National Center on Intensive Intervention (n.d.). *Identifying assessments*. https://intensiveintervention.org/tools-charts/identifying-assessments.

National Center on Student Progress Monitoring. (2006). *Review of progress monitoring tools*. Washington, DC: Author. http://www.studentprogress.org/chart/chart.asp

Neil, A. L., & Christensen, H. (2009). Efficacy and effectiveness of school-based prevention and early intervention programs for anxiety. *Clinical Psychology Review, 29,* 208-215.

Pelham, W. E., Fabiano, G. A., & Massetti, G. M. (2005). Evidence-based assessment of Attention Deficit Hyperactivity Disorder in children and adolescents. *Journal of Clinical and Adolescent Psychology, 3*4, 449-476.

Pett, M. A., Lackey, N. R., & Sullivan, J. J. (2003*). Making sense of factor analysis*. Sage.

Rapport, M. D., Denney, C. B., Chung, K., & Hustace, K. (2001). Internalizing behavior problems and scholastic achievement in children: Cognitive and behavioral pathways as mediators of outcome. *Journal of Clinical Child Psychology, 30*, 536-551.

Revelle, W. (2021). psych: Procedures for Psychological, Psychometric, and Personality Research. R package version 2.1.3 [Computer program]. Retrieved from https://CRAN.R-project.org/package=psych.

Rosseel Y (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, *48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. Psychological methods, 17(3), 354.

Rubin K. H., Chen, X., McDougall, P., Bowker, A., & McKinnon, J. (1995). The Waterloo Longitudinal Project: Predicting adolescent internalizing and externalizing problems from early and mid-childhood. *Developmental Psychopathology, 7,* 751–64.

Rubin, K. H., Coplan, R. J., & Bowker, J. C. (2009). Social withdrawal in childhood. *Annual Review of Psychology, 60,* 141-171.

Sanders, D.E., Merrell, K.W., & Cobb, H.C. (1999). Internalizing symptoms and affect of children with emotional and behavioral disorders: A comparative study with an urban African American sample. *Psychology in the Schools, 36,* 187-197.

Schniering, C. A., Rapee, R. M., Lyneham, H. J., Wuthrich, V., Hudson J. L., & Wignall, A. (2006). *The Cool Kids® Adolescent Anxiety & Depression Program Therapist Manual.* Centre for Emotional Health, Macquarie University.

Shi, D., Lee, T., & Maydeu-Olivares, A. (2019). Understanding the model size effect on SEM fit indices. *Educational and psychological measurement, 79*(2), 310-334. https://doi.org/10.1177/0013164418783530

Smith, S. R. (2007). Making sense of multiple informants in child and adolescent psychopathology: A guide for clinicians. *Journal of Psychoeducational Assessment, 25,* 139-149.

Splett, J. W., Garzona, M., Gibson, N., Wojtalewicz, D., Raborn, A., & Reinke, W. (2019). Teacher recognition, concern, and referral of children's internalizing and externalizing behavior problems. *School Mental Health, 11,* 228-39.

Volpe, R. J., & Gadow, K. D. (2010). Creating abbreviated rating scales to monitor classroom inattention-overactivity, aggression, and peer conflict: Reliability, validity, and treatment sensitivity. *School Psychology Review, 39,* 350-363.

Volpe, R. J., Gadow, K. D., Blom-Hoffman, J., & Feinberg, A. B. (2009). Factor-analytic and individualized approaches to constructing brief measures of ADHD behaviors. *Journal of Emotional and Behavioral Disorders, 17,* 118-128.

Walker, H. M., Nishioka, V. M., Zeller, R., Severson, H. H., & Feil, E. G. (2000). Causal factors and potential solutions for the persistent under- identification of students having emotional or behavioral disorders in the context of schooling. *Assessment for Effective Intervention, 26,* 29–39. doi:10.1177/073724770002600105

Webster-Stratton, C., Reid, M. J., & Hammond, M. (2004). Treating children with early-onset conduct problems: Intervention outcomes for parent, child, and teacher training. *Journal of Clinical Child and Adolescent Psychology*, *33*(1), 105-124.

Volpe, R. J., & Briesch, A. M. (2012). Generalizability and dependability of single-item and multiple-item Direct Behavior Rating scales for engagement and disruptive behavior. *School Psychology Review*, *41*, 246-261.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Volpe, R. J., & Briesch, A. M. (2015). Multi-item Direct Behavior Ratings: Dependability of two levels of assessment specificity. *School Psychology Quarterly, 30,* 431-442. http://dx.doi.org/10.1037/spq0000115

Wickerd, G., & Hulac, D. (2017). Generalizability and dependability of a multi-item direct behavior rating scale in a kindergarten classroom setting. *Journal of Applied School Psychology, 33,* 109-123. http://dx.doi.org/10.1080/15377903.2016.1264530

Table 1

Consumer and Scientific Advisory Panel Ratings

		Consumer Advisory Panel Mean Ratings			Scientific Advisory Panel Mean Ratings			
Item	Item Description[a]	Observable	Malleable	Socially Valid	Construct	Observability	Malleability	Socially Valid
Anxious-Depressed								
Q7	Sad	3.08	2.33	2.83	4.00	3.20	2.60	3.60
Q8	Cries	3.58	3.00	3.08	3.60	3.80	3.00	3.40
Q9	Worries	2.17	2.42	2.92	3.80	2.20	3.20	3.20
Q10	Complains	3.42	2.75	3.00	2.80	3.00	3.20	2.60
Q11	Fearful*	2.50	2.67	2.92	3.80	2.00	3.40	3.60
Q12	Self-conscious	2.25	2.67	2.75	3.00	2.20	2.40	3.00
Q13	Nervous	2.45	2.55	2.73	3.60	2.60	2.80	3.40
Q14	Irritable	2.92	2.42	2.83	3.40	3.20	2.80	3.60
Q15	Pessimistic*	2.42	2.25	2.50	3.00	2.00	2.60	2.60

Socially Withdrawn								
Q1	Relates	2.18	2.27	2.64	3.40	2.80	3.00	3.00
Q2	Shy	2.75	2.17	2.50	3.40	3.60	2.40	3.20
Q3	Quiet	2.73	2.09	2.18	3.20	3.00	2.80	3.00
Q4	Alone	2.73	2.00	2.00	2.40	3.40	3.00	2.60
Q5	Doesn’t join	2.92	2.83	3.17	3.80	3.60	3.40	3.60
Q6	Avoids	2.83	2.58	2.75	3.60	3.60	3.40	3.40

Table 2

Item Descriptives and EFA Loadings

Item[a]	Mean	SD	Skewness	Kurtosis	λ	h^2
Anxious-Depressed Items						
Sad	1.03	1.43	1.51	1.77	.83	.68
Worries	1.62	1.72	0.97	0.08	.77	.60
Nervous	1.29	1.62	1.28	0.88	.76	.58
Self-conscious	1.28	1.52	1.16	0.70	.71	.51
Complains*	0.70	1.27	2.05	3.96	.61	.38
Cries*	1.12	1.65	1.39	0.90	.55	.30
Irritable*	1.30	1.77	1.15	0.10	.45	.20
Socially Withdrawn Items						
Alone	0.86	1.42	1.84	2.79	.91	.83
Does not join	1.14	1.65	1.37	0.86	.84	.70
Relates	1.27	1.71	1.25	0.44	.79	.62
Avoid	0.53	1.11	2.40	5.75	.72	.51
Quiet*	0.75	1.35	2.01	3.54	.43	.18
Shy*	0.77	1.29	2.03	4.03	.41	.17

Note. [a] = abbreviated description of item; * = item deleted. The descriptives were calculated on the full sample, while the factor loadings and the communalities only on the exploratory sample (*n* = 141)

Table 3

Anxious Depressed Scale: CFA Results for One-Factor Model Tested on the Confirmatory Sample (n = 166)

	1	2	3	4	λ	SE	standardized λ
		Observed correlation matrix					
1. Sad	1				1.31	0.16	.79
2. Worries	.71	1			2.06	0.26	.90
3. Nervous	.66	.80	1		1.78	0.24	.87
4. Self-conscious	.71	.74	.73	1	1.56	0.21	.84
		Correlation residuals					
1. Sad	0						
2. Worries	-.01	0					
3. Nervous	-.04	.01	0				
4. Self-conscious	.04	-.02	-.01	0			

Table 4

Socially Withdrawn Scale: CFA Results for One-Factor Model Tested on the Confirmatory Sample (n = 166)

	1	2	3	4	λ	SE	standardized λ
		Observed correlation matrix					
1. Alone	1				1.85	0.25	0.88
2. Does not join	.69	1			1.48	0.21	0.83
3. Relates	.69	.76	1		2.52	0.19	0.84
4. Avoid	.86	.74	.76	1	2.81	0.78	0.94
		Correlation residuals					
1. Alone	0						
2. Does not join	-.04	0					
3. Relates	-.05	.06	0				
4. Avoid	.03	-.05	-.03	0			

Figure 1

Item Characteristic Curves for DBR Anxious Depressed

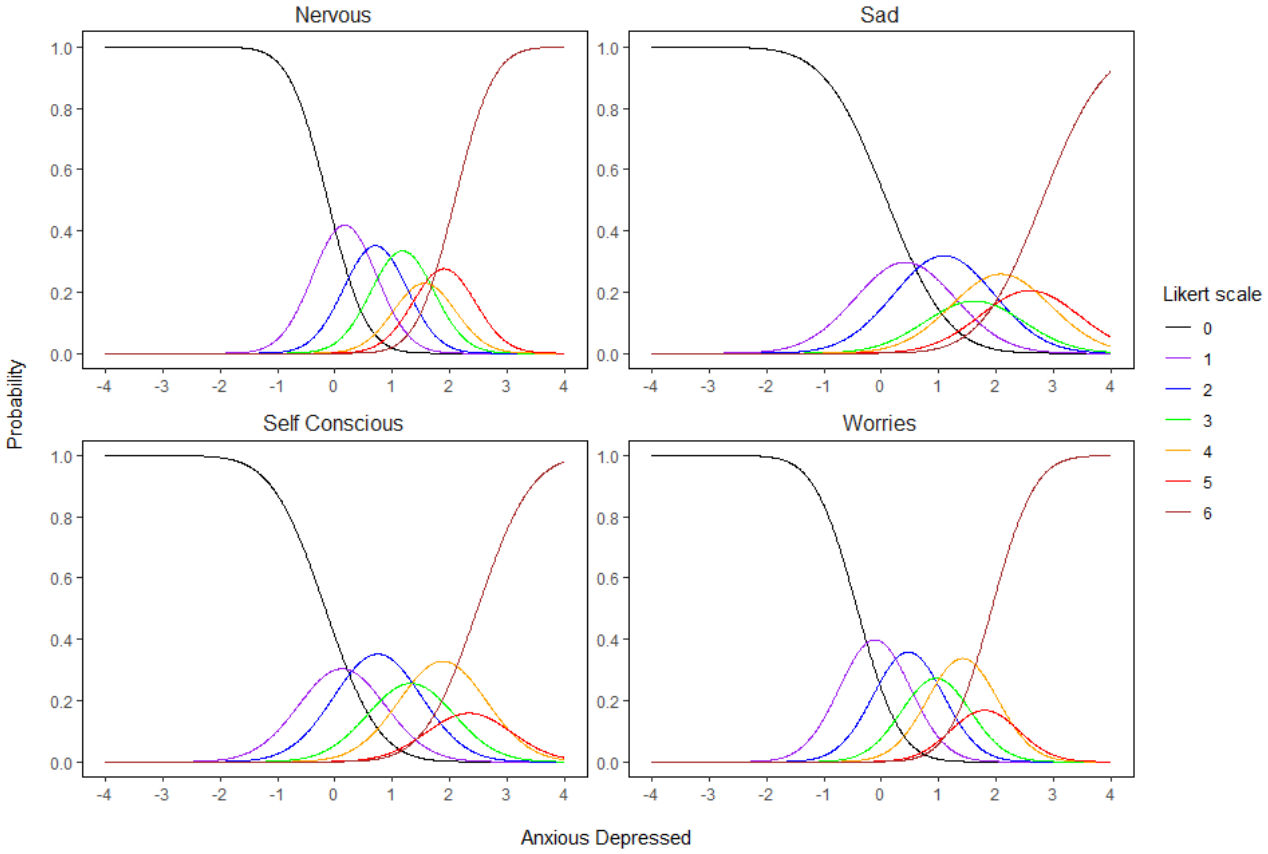

Figure 2

Item Characteristic Curves for DBR Socially Withdrawn